13 Statistical Tools to Improve the Quality of Experiments and Data Analysis for Assessing Non-target Effects

Thomas S. Hoffmeister,[1] Dirk Babendreier[2] and Eric Wajnberg[3]
[1]*Institute of Ecology and Evolutionary Biology, University of Bremen, Leobener Str. NW2, D-28359 Bremen, Germany (email: hoffmeister@uni-bremen.de; fax number: +49-421-218-4504);* [2]*Agroscope FAL Reckenholz, Reckenholzstr. 191, 8046 Zürich, Switzerland (email: dirk.babendreier@fal.admin.ch; fax number: +41-44-377-7201);* [3]*INRA, 400 Route des Chappes, BP 167, 06903 Sophia-Antipolis Cedex, France (email: wajnberg@antibes.inra.fr; fax number: +33-4-92-38-6557)*

Abstract

When testing non-target effects of biological control agents, it is essential that conclusions can be drawn with high precision and confidence. However, testing non-target effects confronts the experimenter with a number of difficulties. First of all, biologically positive cases of not finding any non-target effect are more difficult to substantiate, since in standard statistical hypothesis testing, we can only associate a precise probability to err with rejecting the null hypothesis that assumes no effect, but not with accepting it. The main problem here is the effect size, i.e. the difference from the null hypothesis that is considered biologically meaningful. Secondly, there will usually be a trade-off between the costs associated with increased sample sizes and the confidence of the results of non-target effects testing. Often, sample size will be a limiting factor due to a shortage of animals, space for testing arenas, research funding, etc. Thus, it becomes especially important to optimize the experimental design and to use the most powerful statistical tools to obtain maximum confidence in the test results. Here, we will briefly (i) introduce the reader to common pitfalls of experimental design, (ii) explain the nature of errors in statistical testing, (iii) point towards methods that determine the power of statistical tests, (iv) explain the distribution of the most commonly encountered types of data, and (v) provide an introduction to powerful statistical tests for such data.

Introduction

The last two decades have seen almost a revolution in statistical methods used in ecological investigations, as can be witnessed from a number of recent textbooks on design and statistical approaches in the life sciences (e.g. Crawley, 1993; Hilborn and Mangel, 1997; Crawley, 2002; Grafen and Hails, 2002; Quinn and Keough, 2002; Ruxton and Colegrave, 2003), and from changes in approaches used in more recent

 Environmental Impact of Invertebrates for Biological Control of Arthropods: Methods and Risk Assessment (eds F. Bigler *et al.*)

publications. This reflects both the increased awareness that conclusions in ecological studies need to be drawn in a quantitative manner with high precision and confidence, and that, for a number of reasons, large sample sizes are often difficult to obtain. Thus, the need for powerful statistical tools that allow precise analysis from limited sample sizes is evident. Formerly, the statistical analysis of data in ecological investigations has been fraught with the difficulty that many, if not most, of the data sampled for this purpose are not normally distributed, and are thus not suitable for the parametric 'standard' approaches of Analysis of Variance (ANOVA) and Student *t*-tests. Instead, non-parametric statistics such as, e.g. Kruskal-Wallis and Mann-Whitney *U*-Tests, have been used that are known to be less powerful. In theory, the lack of power of non-parametric statistics may be compensated by larger sample sizes. However, an increase in sample size is often unfeasible for agricultural entomologists, who are usually limited by the time that can be invested, the money that can be spent on experiments, and/or the number of replicates that can be obtained, through a shortage of either experimental fields or insects to work with. Besides such restrictions, several other problems might arise, most of which can be well illustrated by the following example. A couple of years ago, one of the authors of the present chapter heard a talk at an entomological conference, where an investigation into the possible side-effects of genetically modified organisms (GMO) on biodiversity in crop fields was presented. The authors did not find significant treatment effects in most of their tests, but we found it difficult to decide whether the lack of treatment effects was due to a non-optimal experimental design and analysis of the data or whether the conclusion of no effect could be drawn with confidence. Non-target effects of GMOs are an issue of risk assessment that corresponds well with investigations on non-target effects of natural enemies, and thus is used here for an illustration of general problems in design and analysis of risk assessment studies.

This example inspired us to use a computer-generated data set in this chapter to elucidate some of the problems of design and analysis of non-target effect studies, the non-independence of data that leads to pseudoreplicates, the lack of statistical power and the difference between powerful and less powerful statistical techniques. Imagine the following research question and set-up: we wanted to know whether planting genetically modified plants that are resistant to a target pest species would affect the biodiversity of non-target insects in the crop field. For this, we were allowed to do our experiments on a single large field. Imagine further that we partitioned our field into three sections; thus, we had one section with the GMO treatment, adjacent to the section with the conventional crop (serving as control), and on the last section an isoline of the genetically modified crop, which does not express the resistance against the herbivore pest (serving as a second control), was sown. We sampled the biodiversity of non-target insects at ten spots within each of the field sections. Altogether, we received ten data points for each of the three treatments. Imagine we found that the biodiversity of non-target insects in one treatment, e.g. the GMO treatment, was significantly lower. Can we conclude with confidence that the GMO crop affects the biodiversity of non-target insects negatively? Not necessarily. Remember that all the samples for the GMO treatment came from one region of the field. It is possible that the biodiversity of non-target insects had been lower on this side of the field, e.g. due to its proximity to a road. Thus, our spatial clustering of samples has made it impossible to attribute the biodiversity effect to the GMO treatment with confidence, and our ten samples per field section must be considered as being pseudoreplicates.

Now, assume we had chosen to do our experiment in 30 fields, each allotted to one of the three treatments at random, such that we obtained ten fields per treatment. We then find a small trend of decreased biodiversity in the GMO treatment compared with the two control treatments. However, using a Kruskal-Wallis test (because data are

not normally distributed), this trend does not appear to be statistically significant. Can we conclude with confidence that the GMO treatment had no negative effect? To elucidate this, we turned our investigation upside down. Let us assume now that we have an effect of the GMO treatment that reduces the biodiversity by 20%. Using a sample size of ten randomly drawn data points from a Poisson distribution (note that our index of biodiversity is based on species counts, and that counts are usually Poisson distributed), with appropriate means for each of our three treatments, how often would we find a statistically significant difference using a Kruskal-Wallis test? In fact, we would find a significant difference in only about 23 out of 100 cases. Thus, the power of this test is relatively low. Using more powerful statistical tests would increase the power slightly: using a Generalized Linear Model with appropriate Poisson distribution we would find a significant difference in about 27 out of 100 cases.

Even if powerful statistical approaches are employed, the amount of replicates necessary to allow conclusions with high precision can be enormous. In our example given here, 126 instead of 30 fields would have to be studied to detect a reduction of 20% in biodiversity with confidence. In the largest study conducted so far on the side-effects of GMO, a power analysis has suggested that 60 fields per crop had to be sampled across three years to detect effects of ecological significance (Perry *et al.*, 2003; Rothery *et al.*, 2003). An experimental design of this extent will perhaps be impossible in most cases where we wish to test possible non-target effects of biological control agents, and it will not even always be necessary. What will be necessary, instead, is a robust design and the decision by the researchers about what magnitude of an effect is desirable to be detected. This requires knowledge of the power of the statistical testing procedures applied, and in the case of insignificant results, stating the power of the statistical test used. It is only then that we can evaluate whether an insignificant finding is likely to mean that there is no ecological effect, or whether the data are not strong enough to support such a conclusion. This piece of information is still stated only rarely in research papers, and powerful statistics are not yet always employed or even available. Therefore, in the present chapter, we will briefly outline the logic of statistical testing and point towards important advances in statistical techniques for the testing of non-target effects. We will refer to many of the measurement variables mentioned in other chapters of this book and provide suggestions for their analysis. That does not say that we can and do cover everything of importance for the design and analysis of testing non-target effects. However, if we can increase awareness of possible pitfalls of experimental design and point towards solutions or refer to some of the excellent statistics primers, this chapter might help to improve the precision and accuracy of such experiments. Though this chapter focuses on non-target effects of biological control agents, we would further like to stress its relevance for other studies dealing with risk assessment, e.g. non-target effects of pesticides or GMOs.

In the following sections, we will start by reviewing the very basics of statistical testing, i.e. the hypotheses involved in statistical testing and the errors associated with accepting or rejecting those hypotheses. Subsequently, we will discuss the effect size and power of statistical tests, measurements that are of high relevance given a statistical test does not return significant results. Further, the need to obtain independent data for statistical testing and the danger of pseudoreplication will be explained, and also how randomization can prevent pseudoreplication. Building upon this, we present powerful statistical tools, such as Generalized Linear Models and Cox regressions, for the analysis of the kind of data that will typically be generated when assessing non-target effects.

Two Ways to Err in Statistical Testing (α- and β-errors)

By performing an experiment it remains impossible to prove, for example, that a nat-

ural enemy will never attack a non-target host or prey. Using a sound experimental design, we can aim only at achieving high accuracy and precision in what we conclude from the sample that we have tested. Yet, using standard statistical procedures, there is always some possibility that our interpretation of the data is wrong. This is due to the fact that all the measurement variables we are interested in are usually subject to random variation (i.e. variation between sample units that cannot account for a treatment factor under consideration), and that our conclusion is based on a sample rather than on the entire population. Since we conclude from a statistical test either that the null hypothesis (H_0) is wrong, and can thus be rejected, or that the alternative hypothesis (H_1) is wrong, and thus H_0 cannot be rejected, we have two ways to err (Table 13.1). An α-error (also called Type I error) occurs if our experimental results suggest there is an effect of the factor of interest on the variable we wish to explain (the so-called 'dependent variable') when in fact there is none, thus if we reject H_0 provided that H_0 is correct. A β-error (also called Type II error) occurs if there is a true effect of the factor in question, but our experiment fails to detect this effect, thus if we do not accept H_1 when H_0 is wrong (Fig. 13.1, Table 13.1). Only the α-error can be immediately quantified: the *P*-value associated with a test statistic immediately provides the probability of committing an α-error. Usually, the null hypothesis is rejected if the probability of committing an α-error is 0.05 or less. In that case, the alternative hypothesis is accepted.

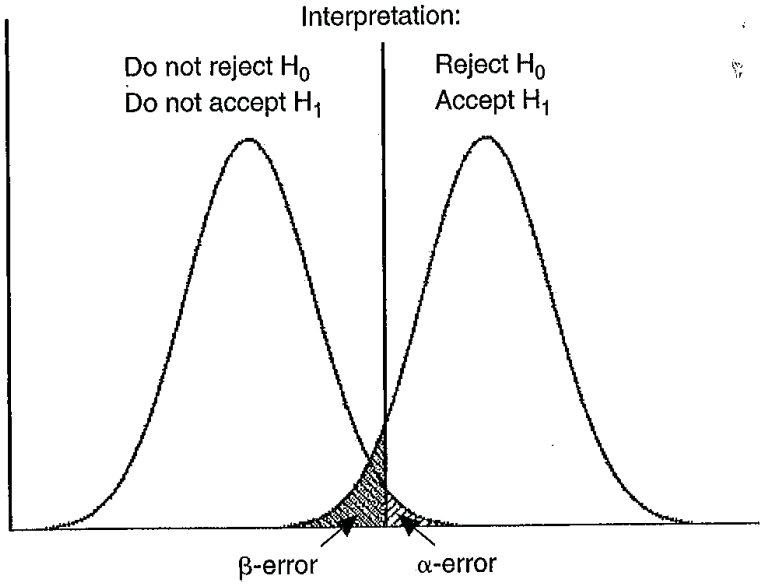

Fig. 13.1. Graphical representation of α-error (area hatched in white and black) and β-error (area hatched in grey and black) probabilities, using a one-sided *t*-test, comparing, e.g. encounter rates of biological control agents with non-target hosts. The curves on the left (for the null hypothesis) and right (for a specified alternative hypothesis) represent the probability sampling distribution of the statistical test done. Note that, usually, the alternative hypothesis is not specified, i.e. H_1 is just different from H_0, and the probability distribution of the statistical test done for H_1 is unknown (modified from Quinn and Keough, 2002).

Table 13.1. Hypothesis testing: the truth associated with a decision derived from a statistical test when the null hypothesis is in fact true or not true.

		Decision	
		H_0 is not rejected	H_0 is rejected
Truth according to model	H_0 true	correct	α-error (type I)
	H_0 not true	β-error (type II)	correct

It should be noted, however, that while the statistical test returns a precise error probability for rejecting H_0, it is not possible to associate a precise error probability with accepting H_1 (see Fig. 13.1, where the α-error is associated with the probability distribution of H_0 and not with H_1). Moreover, it is important to mention that the α-level is equal to the *P*-value of the test only if we perform a single test on a given set of data. If we wish to perform multiple pairwise comparisons between, e.g. means from an experiment with more than two treatments, the probability of making at least one α-error by chance among those tests increases with the number of tests performed. This probability of making one or more α-error is called the family-wise α-error rate. When such tests are not independent from each other, e.g. if one data set is used more than once in a test, the family-wise α-error rate becomes difficult, if not impossible to calculate precisely. Yet, several procedures have been put forward to correct for multiple testing. The best known is the Bonferroni procedure, where the α-error is divided by the number of tests performed to obtain a new significance threshold and to keep a global α-error for the whole testing procedure. However, this procedure is overly conservative, i.e. in danger of committing β-errors (to elucidate this, imagine shifting the border between accepting and rejecting H_0 in Fig. 13.1 to the right; while α-error decreases, β-error increases). The standard procedure for correcting for multiple testing is the sequential Bonferroni procedure suggested by Holm (1979), where *P*-values of all m tests are ranked from largest to smallest: the smallest *P*-value is tested at α/m, the second smallest is tested at $\alpha/(m-1)$ and so on, until the first non-significant result occurs. Recently, this procedure has also been criticized for being too conservative (Moran, 2003), and there is an ongoing discussion about the optimal way to correct for multiple testing (Garcia, 2004; Neuhäuser, 2004; Verhoeven *et al.*, 2005). For a good overview on this topic, we recommend the reader consults Quinn and Keough (2002). An important aspect that needs particular attention when testing for non-target effects, if we want to err on the side of caution, is that it might be more important to know the probability that an effect actually exists, given we did not find an effect (the β-error), than accurately quantifying the α-error. An α-error fixed at 0.05 is not necessarily meaningful. What we need to know instead is the power of the statistical test (see next section, below), which might lead us even to compromise between α- and β-errors (see below).

Example

Taking one of our above-mentioned data sets about the effects of GM-plants on the biodiversity of non-target insects, our null hypothesis would be that in plots with all three treatments (GMO, non-GM isoline and conventional crop) the insect biodiversity would be the same. Now, we will not use a Kruskal Wallis test (K-W-test) as in the introduction, because it would not be easy to calculate the β-error associated with the K-W-test. Instead, by using an ANOVA on square-root transformed data (to achieve Gaussian distribution of data), we find that the α-error is $P = 0.584$. Thus, rejecting the null hypothesis and accepting that there is an effect of plant treatment on biodiversity, one would err in 58.4% of the cases. Using a programme for Power analysis (see below) one can calculate the β-error. In our case, the β-error is 0.768, if we wish to be able to detect a 20% difference in biodiversity of non-target insects. Thus, by not rejecting the null hypothesis, and consequently, by not accepting the alternative hypothesis, one would err in 76.8% of the cases. Obviously, this data set is insufficient for either accepting the alternative hypothesis or for not rejecting the null hypothesis with confidence.

Ecological Effect Size, Replicate Number and the Power of Statistical Tests

Statistical power is the probability that a given test will result in rejection of the null hypothesis when that null hypothesis is,

indeed, false. Hence, power = $1-\beta$. For any particular test, power is dependent on the α-level, the sample size, the sampling variance and the so-called 'effect size' (ES). The ES can be regarded as the magnitude of the departure from the null hypothesis (observed ES), or as the difference between the values considered in the null and the alternative hypotheses (see Fig. 13.1 and below).

There are two general approaches in Power Analysis (PA). The first one is a priori PA, where one aims to estimate the number of replicates necessary to reach a given power in an experiment. This can be done by specifying the effect size, the α-level, the desired power and (dependent on the type of analysis) the standard deviation, which has to be estimated from preliminary experiments or from the literature. It should be stressed, however, that estimates for the assumed variance of the data are crucial. Carey and Keough (2002) have shown that the calculated sample size can vary by an order of magnitude depending on what dataset was used as a baseline for variance. The second approach is a post hoc analysis, where the researcher calculates the power achieved in an experiment where the null hypothesis could not be rejected. While general agreement exists on the importance of a priori PA, there is considerable debate on the value of post hoc PA. In particular, parameters are estimated based on the sample data in post hoc PA and are therefore interdependent. Since these estimates are subject to sampling error, the computed values for power are also subject to error and thus should be viewed with some caution.

Obviously, the statistical ability to detect an effect (i.e. the power) increases with the size of that effect and, in fact, power is extremely sensitive to one's choice of effect size (Cohen, 1988). There are several approaches for calculating post hoc power, and the effect size plays a crucial role in all of them. The first approach is to use the observed effect size, e.g. taking the difference between the control and the treatment from the data, and variance. However, this has clear flaws which form the basis of large parts of the criticism of post hoc PA (Hoenig and Heisey, 2001; Di Stefano, 2003). Actually, the *P*-value and power are dependent on the observed effect size such that tests with high *P*-values tend to have lower power, and vice versa. Therefore, calculating power based on observed effect size and variance adds no new information to the analysis (Thomas, 1997).

The second approach is to use a predefined effect size and observed variance. Although it can be often difficult to define effect size properly, a useful approach, especially in the context of assessing non-target effects, has been to estimate an effect that can be considered biologically significant. For instance, if an earlier study showed that 40% mortality caused populations to decrease in a wider context, this figure could be used as effect size for another study. As was shown in detail and exemplified with an example by Thomas (1997), this second approach appears valuable and allows one to evaluate whether the sample size and α-level were likely to result in detection of a biologically meaningful effect.

A third approach is to establish an effect size based on the null and the alternative hypotheses. However, in this case the latter needs to be formulated quantitatively, which is only possible in certain instances. In the absence of any strong arguments that are independent of the hypothesis being tested, the selection of an effect size becomes arbitrary. However, in the case that effect size could neither be calculated based on biological significance nor from the alternative hypothesis, some conventions can be used that were established by Cohen (1998). He suggested using large, medium or small effects as a convention, but the exact size of these effects depends on the type of statistical analysis used. Many software packages readily provide the standardized effect, which is basically the difference between H_0 and H_1 divided by the standard deviation of the data. Although this avoids specifying the sampling variance, we feel it unwise to use the standardized effect, because it is poorly related to any biologically meaningful

effect. Rather, we recommend calculating effect size based on either biological significance or on a quantitative alternative hypothesis, but we also believe that it is useful to put the ES of a study into context and to compare it to the procedure proposed by Cohen (1998). As a consequence of the importance of the ES outlined above, we also recommend strongly to report in detail on the ES underlying the analysis, rather than giving only a figure for β or power (cf. Steidl *et al.*, 1997).

As a special case of PA, the maximum detectable effect size could be calculated; this can be performed easily by fixing the power and the α-level appropriately. For instance, a researcher might wish to know about the effects he/she would have been able to detect given that the power is 0.8, a figure that has often been used. What constitutes a sufficient power is not absolutely fixed, though conventions of 0.8 or 0.95 have been suggested in the literature as high power (Cohen, 1988). However, in studies on environmental impact it is debatable why one should be satisfied with accepting a four-times higher β- than α-error, which is the case when using the 0.8 value. In contrast, one would like to be at least as confident in avoiding β-errors and α-errors alike in such investigations. Thus, a researcher conducting experiments on potential non-target effects of a biological control agent could ask what maximum possible effect size is consistent with $\alpha = \beta$. In this context, it is important to note that in studies dealing with non-target effects, it may be reasonable to increase the α-level, thereby increasing power. Eventually, it depends on the costs associated with specific non-target effects. If the costs of committing β-errors are especially high, PA allows one to adjust α/β to reflect those costs (Rotenberry and Wiens, 1985).

As an alternative to classical PA, the application of confidence intervals and equivalence testing has been suggested recently (Hoenig and Heisey, 2001; Andow, 2003). Demonstrating such equivalence requires reversing the traditional burden of proof. In equivalence testing, the null hypothesis states that a large effect exists in either direction, i.e. the actual treatment effect (D) is larger than a predefined δ (H_0:$|D| > \delta$). The alternative hypothesis is the hypothesis of equivalence, or H_1:$|D| = \delta$. Again, this kind of analysis depends on the knowledge of what a large (biologically meaningful) effect is, and the determination of delta is similarly as difficult as determination of the effect size, as discussed above. Given the large uncertainty in this area, it is difficult to give advice on this, though the general idea is appealing for decision-makers in risk assessment (Peterman, 1990).

In conclusion, a priori PA can be a valuable aid in the design of any study and, in particular, for monitoring programmes (see Barratt *et al.*, Chapter 10, this volume). In addition to the information on sample size necessary to detect a given effect, it is also very valuable for reducing the cost of large-scale programmes as far as possible. Depending on the research question, post hoc PA also can be very useful, particularly because it is not always possible to conduct an ideally high number of replicates. It should be stressed that it is not possible with PA to associate an unambiguous probability of being correct in not rejecting the null hypothesis although, unfortunately, this has been done quite often in the past (see Peterman, 1990). Instead, it is only possible to argue that, with a probability of $(1-\beta)$, there is no difference from the H_0 greater than the effect size. If both the ES and β are small (and consequently the power is high), it is reasonable to conclude that the effect is negligible. It is particularly important in studies on non-target effects that a conclusion from a non-significant statistical result should be subject to the same stringent probability standards as a positive conclusion from a significant statistical result. Power analysis could be used to provide these standards.

Programs available

A comprehensive review on this topic was written by Thomas and Krebs (1997), and we do not attempt to provide a similar

detailed compilation here. Instead, we would like to refer to some published information – also on the internet – and highlight a few recent developments. Since the influential paper by Thomas and Krebs (1997), some significant advances have been made, wherein some programs are able to calculate the power for regressions, comparisons of means (ANOVA and General Linear Models) or proportions (χ^2 tests), for correlation tests and survival analysis. However, there are still several statistical tests for which PA is not available and, unfortunately, this includes the Generalized Linear Models, which can be a very powerful statistical tool for data that do not follow a Gaussian distribution. There are also possibilities for calculating power for other tests, but efforts to do this can vary from relatively simple to challenging. For instance, Monte Carlo simulations can be used to calculate power for non-parametric tests (Peterman, 1990). Alternatively, data have to be transformed to fit the assumptions of tests that allow PA, e.g. log-transformation or square-root transformation for count data, arc sine square-root transformation for proportions (see, e.g. Quinn and Keough, 2002, or another standard statistics textbook, for further information). Information on programs and their strengths and weaknesses can be also obtained from the following homepages: List of programs (from 1996) (http://www.insp.mx/dinf/stat_list.html) and paper by Thomas and Krebs (1997), (http://www.zoology.ubc.ca/~krebs/power.html).

Examples

Let us, again, take a look at the example data provided in the introduction. Using ten fields for each treatment, the effect of GM plants on insect biodiversity was tested. If we were to analyse those data with ANOVA, we would have to transform the species numbers to receive data with Gaussian distribution. Square root transformation ($y'=\sqrt{y+1}$) could be favourable in our case. If our control plots could harbour eight non-target species and we wish to be able to detect a 20% loss of biodiversity (i.e. 6.4 species on average), the resulting transformed means for species numbers would be 3, 3, and 2.72 for the three treatments, respectively, and the standard deviation would be approximately 0.5 for all treatments. A simple ANOVA did not detect a significant effect. Entering the above-mentioned values in a programme for PA returns an effect size of ES = 0.2828, and thus what is conventionally described as medium effect size. With a total sample size of 30 the power is $(1-\beta) = 0.2397$. How many replicates would be needed to achieve a power of 0.8 with such an effect size? Using an a priori test in the programme for PA we receive a necessary sample size of n = 126. Thus, to demonstrate with high confidence that no effect exists would require a much larger study (see, e.g. Lang, 2004 for an estimate of necessary sample sizes for non-target effects of *Bt*-plants).

Using another example, let us see how large the sample size should be in a non-target effects study of an insect natural enemy. Using the above-mentioned example of Thomas (1997), where the non-target population would be affected only if the mortality were higher than 40%, we can use 0.4 as effect size in an a priori test. If we were to achieve a power of 0.8, the necessary sample size in an experiment with two treatments would be n = 52.

Avoid Being Trapped in Pseudoreplication

In a seminal paper, Hurlbert (1984) published a review with respect to proper replication of 176 field experiments covering 156 papers published in ecological journals between 1960 and 1983. Disturbingly, he found that of the 101 studies applying inferential statistics, 48% contained pseudoreplication. Pseudoreplication occurs whenever 'inferential statistics are used to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent' (Hurlbert,

1984). Statistical independence means that each individual data point might positively or negatively deviate from the population average due to random variation not related to the deviation of another point. An example of lack of statistical independence is given in the introduction, where samples of a study on effects of GMOs on biodiversity were segregated by treatment and, thus, differences attributed towards the treatment could equally well have been attributed to some factor typical for the section of the field the samples came from. In this case, the effects of treatments are potentially confounded with inherent differences between field plots. Although the awareness of researchers of avoiding pseudoreplication has increased and fewer studies contain analyses with pseudoreplicated samples, Heffner *et al.* (1996) and Ramirez *et al.* (2000) found, in a recent study on pseudoreplication in experiments on the olfactory response of insects, that an alarming 46% of 105 studies were pseudoreplicated, because of either a lack of independence in the stimulus or the experimental device, the repeated use of experimental animals or the use of groups of animals.

Thus, pseudoreplication is still an issue in the design of experiments, and much care has to be taken to avoid any spatial or temporal segregation of samples from different treatments. For example, when testing the host specificity of biological control agents, it is essential that insects for the tests on non-target hosts do not come from one rearing container or incubator and control animals (for the test on target hosts) come from another, or that non-target hosts are always tested in the same container or field cage or on the same plant while target hosts are tested in another cage or on another plant. Equally, positions of experimental units within an experimental chamber or on a field plot need to be switched between treatments to avoid confounding effects of differences in temperature and light conditions, etc. In the same manner, the full set of trials on non-target hosts should not be conducted before tests with target hosts are carried out. Randomization of testing order, or random assignment to plants or test cages, ensures that pseudoreplication can be avoided. For further reading, we encourage the reader to take a look at the section on pseudoreplication in Ruxton and Colegrave (2003).

Experimental Design: is Randomization Feasible?

Basic textbooks on statistics always stress the point that, in order to draw relevant conclusions from an experiment, all treatments, replicates, etc., should be randomized. But what does that mean? Randomization is a process that assigns each replicate of each measured unit (animal, field, species, etc.) to each treatment in a random order, rather than by choice. By doing this, any effect observed will be unequivocally attributed to the treatment studied, and not to lurking variables or uncontrolled factors which might vary over the length of the experiment. For example, if one was interested in estimating the host-range specificity of different potential biological control agents for a pre-release evaluation of non-target risks, he/she would sequentially offer several potential host species to the different biological control agents studied (see van Lenteren *et al.*, Chapter 3, this volume for a detailed description of the proposed method to be used). In this case, it would be preferable to: (i) test the different host species in a random order for the different biological control agents, and (ii) test each host species, with the different biological control agents taken in random order as well. Indeed, in the case where the different host species are always tested in the same order, uncontrolled factors varying with the duration of the experiment could influence the results and lead to differences that might be wrongly interpreted as being due to differences between species. Also, if all potential host species are tested successively on each biological control species, a difference observed between biological control species might simply be due to uncontrolled factors varying with the total duration of the experiment.

The goal of randomization is to produce comparable groups of replicates in terms of general animal, field, etc., characteristics and other key factors that might affect the outcome of the result obtained. In this way, all groups of replicates are as similar as possible at the start of the study. At the end of the study, if group outcomes differ between each other, the investigators can conclude with some confidence that the treatment tested really influenced the results obtained.

Most of the time, randomization is performed by means of a computer program, coin flips or a table of random numbers to assign each measured unit to a particular treatment. Advanced additional methods are sometimes used.

Is randomization always feasible, especially in evaluating non-target risk in biological control programmes? Unfortunately, the answer is likely to be 'no'. In the example given above, where we wanted to estimate the host-range specificity of different potential biological control agents, it would probably be unrealistic to design an experiment in which all host species tested and all potential biological control agents compared were randomized. Regarding the fact that the experimental scheme is based on a succession of different measures (see van Lenteren *et al.*, Chapter 3, this volume), having everything randomized would indeed imply having available, during the total duration of the experiment, a sufficient number of all host and biological control agent species at the right stage. In most cases this would simply be not feasible for economic or spatial reasons. All of this should be kept in mind and, if real randomization appears not feasible, results of the experiments should thus be interpreted with caution.

A Unified Approach Instead of a Menu of Tests, General and Generalized Linear Models

When the traits to be analysed follow a Gaussian (also called 'Normal') distribution, standard *t*-tests, ANOVA or regression analyses can be used to statistically test the effect of a treatment. All these different 'classical' methods assume that the distribution of residuals around the fitted model (i.e. the error distribution) is normal (Gaussian). These different methods, which most readers will be familiar with, are called 'General Linear Models', since in its simplest form, a linear model specifies the (linear) relationship between the variable (or response) y, to be explained (the so-called 'dependent' variable), and a set of predictors, independent variables, the xs, such that

$$E(y) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k \quad (1)$$

In this equation, b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables x_1 to x_k) computed from the data. So, for example, one could estimate (i.e. predict) the weight of a parasitoid female as a function of the type and number of hosts it feeds on. For many data analysis problems, estimates of the linear relationships between variables are adequate to describe the observed data, and to make reasonable predictions for new observations. However, as we have seen previously (see Box 13.1), most of the biological traits that have to be measured to estimate non-target risks of biological control agents do not necessarily follow a Gaussian distribution. In such cases, the relationship between the variable (or response) y to be explained cannot adequately be summarized by a simple linear equation, for two major reasons:

DISTRIBUTION OF THE DEPENDENT VARIABLE. First, the dependent variable of interest may have a non-continuous distribution and, thus, the predicted values of the statistical model should also follow the respective distribution. Any other predicted values are not logically possible. For example, an investigator may be interested in predicting one of two possible discrete outcomes (e.g. a host is accepted or not). In that case, the dependent variable can take on only two distinct values, and the distribution of the dependent variable is said to be binomial. Another example would be to predict how

Box 13.1. Measurement variables and their distribution

Many 'classical' statistical approaches rely upon the assumption that the probability distribution of data from samples and the error terms of the statistical analyses (the residuals) are distributed normally, i.e. Gaussian. With many of the measurement variables we collect in non-target testing of biological control agents, these assumptions are not met. Count data such as, e.g. number of mature eggs of a female, are usually Poisson distributed, data for percentages are Binomial, and data for longevity are usually Exponential or sometimes Gamma distributed. In theory, it is possible to transform many kinds of data such that the assumptions of parametric tests are met, and those tests are also robust against small deviations from the assumptions; but first of all it is hard to estimate the extent of the robustness against deviations from normality in error terms and, secondly, it is often advisable to use actual data rather than transformed data to meet assumptions. The probable most commonly collected types of data are listed in the table below. Note that deviations from the distributions mentioned in the table might occur in individual cases and that, in general in statistical testing, residuals should always be inspected for the adequacy of the model.

Measurement variables often found in non-target testing of biological control agents and their distribution.

Measurement variable	Distribution (most likely)
attack rate (per unit time)	Gaussian
dispersal capacity	Gaussian, or Poisson if counts
diurnal periodicity	Gaussian
egg load	Poisson if counts or Binomial if proportion
encounter rate (per unit time)	Gaussian
fecundity	Poisson if counts or Binomial if proportion
frequency of mating	Poisson if counts or Binomial if proportion
growth rate	Gaussian
host acceptance	Binomial
insertion/deletion of genes	Poisson
latency to attack	Gamma
morphology	Gaussian
rate of development	Gaussian
rate of predation/parasitism	Binomial
spatial distribution (i.e. counts)	Poisson or Negative binomial
survivorship/mortality	Gamma
thermal budget (degree-days)	Gaussian

many females a male can mate with. If we were to study actual numbers and not average number of matings per male, the dependent variable (i.e. number of females mated) is discrete (i.e. a male can mate with one, two or three females and so on, but cannot mate with 3.46 females or with fewer than 0 females), and most likely the distribution of that variable is highly skewed (i.e. most males will mate with one, two or three females, fewer will mate with four or five, very few will mate with six or seven, and so on). In this case it would be reasonable to assume that the dependent variable follows a so-called Poisson distribution.

LINK FUNCTION. A second reason why a simple linear model might be inadequate to describe a particular relationship is that the effect of the predictors on the dependent variable may not be linear in nature. For example, the relationship between the fecundity of a synovigenic parasitoid female and its age is most likely not linear in nature. Under standardized conditions, fecundity will not markedly differ between females of one or two days of age, whereas such a difference will probably be greater between older females, even with only one day's age difference. Probably some kind of a power function would be adequate to

describe the relationship between females' age and fecundity, so that each increment in days of age at older ages will have greater impact on females' fecundity, as compared to each increment in days of age during early adult life. Put in other words, the link between age and fecundity is best described as non-linear, or rather as a power relationship in this particular example.

Generalized Linear Models are a generalization of general linear models and can be used to predict responses both for dependent variables that are not normally distributed and for dependent variables which are non-linearly related to the predictors. Actually, general linear models can be considered as special cases of the generalized linear models. In general, in linear models, the dependent variable values have a normal distribution and the link function, which 'connects' the dependent variable to a linear combination of predictor variables, is a simple identity function (i.e. the linear combination of values for the predictor variables is not transformed).

To illustrate this, equation (1) gave the general linear model linearly associating a response variable y with values on the x variables, while the relationship in the generalized linear model is assumed to be

$$E(y) = g(b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k) \quad (2)$$

where g(...) is a function. Formally, the inverse function of g(...), say f(...), is called the link function, so that

$$f(E(y)) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k \quad (3)$$

where E(y) stands for the expected value of y.

Various link functions (see McCullagh and Nelder, 1989) can be chosen, depending on the assumed distribution of the y variable values. Table 13.2 gives the four main Generalized Linear Models that can be used in experiments performed to estimate non-target risks of biological control agents.

The values of the regression parameters (and their variance and covariance) in the Generalized Linear Model are obtained by a so-called maximum likelihood estimation, which requires iterative computational procedures. Several statistics packages are currently available for doing this. Then, tests of the significance of the effects in the model can be performed via the Wald statistic, the likelihood ratio or score statistic. Detailed descriptions of these tests can be found in McCullagh and Nelder (1989).

In summary, Generalized Linear Models are powerful and efficient tools for analysing the sort of data collected in experiments performed to estimate non-target risks of biological control agents. Just a brief overview has been provided here, and there are several textbooks that provide a thorough description of this sort of statistical modelling approach (e.g. Hosmer and Lemeshow, 1989; McCullagh and Nelder, 1989). We strongly recommend readers of this chapter to consult them.

Examples

Using again our example from the introduction, we may analyse one of our computer-

Table 13.2. List of the main Generalized Linear Models that can be used in experiments performed to estimate non-target risk of biological control agents. Link functions indicated are the most 'popular' ones. Others can be used in particular cases (see McCullagh and Nelder, 1989 for an exhaustive description).

Distribution	Model description	Appropriate link function	Type of data analysed
Normal	Traditional linear model	identity: $f(y) = y$	Normally distributed traits
Binomial	Logistic regression	logit: $f(y) = \log\{y/(1-y)\}$	Fractions (proportions)
Poisson	Log-linear model	log: $f(y) = \log(y)$	Counts
Gamma	Gamma model with inverse link	inverse: $f(y) = 1/y$	Time durations

generated data sets using a Generalized Linear Model. Since we count the number of species in each field plot, our data are most likely Poisson distributed. Specifying a Generalized Linear Model with Poisson distribution and log link function, and using the number of species per plot as response variable and the crop treatment (GM-plants, non-GM isoline and conventional crop) as factor, we find a *P*-value of 0.0962; thus, there is an insignificant trend in the data (Fig. 13.2a). An analysis of these data using an ANOVA on square root-transformed data yields a *P*-value of 0.147. A visual comparison (Fig. 13.2b and c) and statistical tests of the normality of the standardized residuals from both analyses (P = 0.515 and P = 0.474, respectively) suggest that the Generalized Linear Model is the slightly more adequate approach to analyse these data. Note that in both cases the statistical result is insignificant and, thus, the null hypothesis of no effect cannot be rejected, but also that the power analysis suggests a lack of power to conclude with confidence that there is no effect.

As a second example, imagine a large arena choice test as suggested by van Lenteren *et al.* (Chapter 3, this volume). We have three different treatments, with ten field cages each: (1) with the target host (or prey, which is used synonymously here) and non-target host present in the same field cage together with the natural enemy, (2) with only the non-target host and the natural enemy in the same field cage, and (3) with only the target host and the natural enemy in the same field cage. We are interested in whether the target host is killed at a higher rate than the non-target host and whether the mortality of the non-target host depends upon the fact of whether the target host is available to the natural enemy or not. We will not test whether the mortality rates of target and non-target host are equal within treatment (1), because these data would not be independent. Rather, we will test whether the mortality of non-target hosts in treatment (1) is equal to the mortality of non-target hosts in treatment (2) and equal to that of the target hosts in treatment (3) (this is our null hypothesis). Again, we will use computer-generated data. Given that the mortality rates found were 4.1%, 10.6% and 50.5% in (1), (2) and (3), respectively, we use a Generalized Linear Model with binomial distribution and logit link and find a significant effect overall and also between treatments (Table 13.3). Thus, in this example, the non-target host is attacked at a relatively low rate, and even less so when target hosts are available. This result is visible from the estimates in Table 13.3, where the estimate for mortality is positive and thus higher in treatment (2) than in treatment (1), and much higher (more than three times higher) in treatment (3) than in treatment (1).

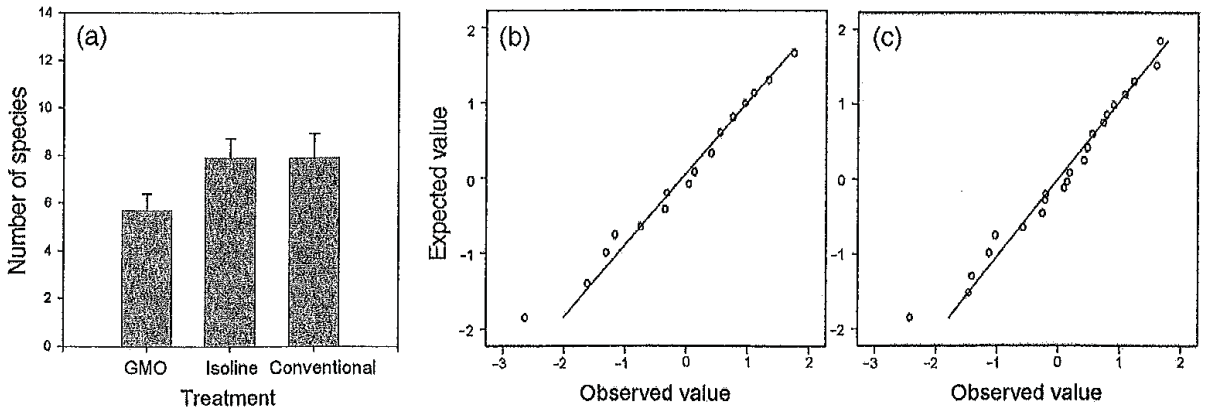

Fig. 13.2. Simulated average (+ SE) effect of plant treatment on non-target insect species (panel a). The computer-generated data were analysed by means of a Generalized Linear Model with a log link function and ANOVA on square root transformed values, respectively. Panels (b) and (c) show Normality Plots for the standardized residuals of the respective tests. The relationship in panel (b) shows a slightly better fit with normality assumptions than in panel (c).

Table 13.3. Results of a Generalized Linear Model on computer-generated data for the mortality rates of target and non-target hosts in large arena choice tests, using an experimental set-up as suggested by van Lenteren *et al.* (Chapter 3, this volume) (for details, see text).

Parameter	Treatment	Estimate	DF	χ-Square	Pr > ChiSq
Intercept		−3.2591	1	378.47	<0.0001
Target host	(3)	3.2511	1	329.63	<0.0001
Non-target host in no-choice test	(2)	1.0839	1	30.14	<0.0001
Non-target host in choice test*	(1)	0	0	0.0000	

* In the statistics package SAS, which was used here, the last treatment (in this case (1)) is set to zero by convention and the difference between the last and all other treatments (2) and (3) is tested.

Repeated Measurements in Generalized Linear Models

Sometimes, the same individual insect or the same experimental plot is systematically sampled more than once in the course of an experiment. Data from such samples violate the assumption of the independence of data points since they do not have an equal probability of deviating positively or negatively from the population average, but contain some variation due to inherent properties of the individual animal or experimental plot. They can thus be considered pseudoreplicates that cannot be entered into statistical tests as independent data points. Liang and Zeger (1986) introduced Generalized Estimating Equations (GEE) to Generalized Linear Models as a method of dealing with such correlated data. GEE is not available in all statistical packages that provide Generalized Linear Models, but at least SAS (procedure Genmod) and S-plus/R provide GEE. They require that a variable identifies the repeated subject and that the model statement refers to this variable as repeated. More details about GEE can be found, e.g. in Quinn and Keough (2002).

Example

Imagine the following field experiment (see van Lenteren *et al.*, Chapter 3, this volume for the rationale of a field test on non-target effects of a biological control agent): we wish to monitor the mortality induced by the natural enemy on the target and non-target hosts across a time period after the release of the natural enemy. We are especially interested in whether the attack rate on non-target hosts depends upon the density of the target host, which may decrease over the course of the experiment. Again, we will use computer-generated data. In our computer program, we select ten different field plots that we resample at five different times. Over time, the number of target hosts per field plot decreases while the mortality of the non-target hosts increases (Table 13.4). However, in order to

Table 13.4. Computer-generated data for a field test on non-target effects as a function of time (sampling date) and density of target hosts. Means and standard errors of ten field plots.

Sample	Density of target host	Mortality of non-target host
1	996.6 ± 10.9	1 ± 0.4294
2	493.4 ± 7.86	38 ± 0.516
3	289.9 ± 5.78	5.6 ± 0.872
4	172.0 ± 2.78	7.3 ± 0.870
5	102.3 ± 3.65	11.8 ± 1.572

elucidate the effect of target host density, we enter sampling date and density of the target host as covariates in the model. The Generalized Linear Model allows us to separate the effect of sampling times and target host density. The GEE model for repeated measurements takes care of the fact that we resample the same field plots, and thus target host densities and the mortality rate of the non-target hosts in each plot are not independent. With both variables, sampling date and the density of the target host, in the model we do not find a significant effect (Table 13.5). However, by removing the variable with the least explanatory power from the model (i.e. sampling date), we find that the density of the target host affects the mortality rate of the non-target host (Table 13.5). Estimates from the model show that mortality of non-target hosts increases with decreasing density of the target host, indicating a switch of the natural enemy to a non-preferred host when the preferred host is less available.

Time as a Measurement Variable: Cox Regression and Survival Analysis

To estimate the potential impact of natural enemies on their host and potential non-target host populations, it is often useful to acquire knowledge about the survival times of such insects. Survival data of insects are not normally distributed, but rather the probability λ of an insect being dead at time t, in the simplest case, can be considered to be constant. This leads to an exponential distribution of the data with mean survival time $1/\lambda$, well known from the decay of radioactive particles and a series of population dynamics models, e.g. Ricker fishery models. Here, the arithmetic mean survival time is a poor predictor of the longevity and, usually, the median is used. Besides considering an exponential distribution of the survival times, predictors of a generalized linear model with the more general Gamma distribution and inverse link function give, as this was stated in the previous

Table 13.5. Analysis of Generalized Estimating Equations (GEE) parameter estimates of a Generalized Linear Model for repeated measurements of the mortality rate of non-target hosts in a field test (data from Table 13.4). The upper part of the table shows the analysis with both sample date and density of target host as explanatory variable, which results in an insignificant model. Removing the variable with least significance (i.e. 'sample date') leads to a model that demonstrates a significant and negative relationship between the density of the target host and the mortality of the non-target host (lower part of the table).

	Empirical Standard Error Estimates			
Parameter	Estimate	Standard Error	Z of Wald test	Pr > |Z|
Intercept	−3.2420	0.9599	−3.38	0.0007
Sample date	0.2715	0.1917	1.42	0.1568
Density target host	−0.0016	0.0011	−1.44	0.1503

	Score Statistics For Type 3 GEE Analysis		
Source	DF	Chi-Square	Pr > ChiSq
Sample date	1	1.44	0.2304
Density target host	1	2.12	0.1451

	Empirical Standard Error Estimates			
Parameter	Estimate	Standard Error	Z of Wald test	Pr > |Z|
Intercept	−1.8476	0.1351	−13.68	<0.0001
Density target host	−0.0031	0.0005	−5.99	<0.0001

	Score Statistics for Type 3 GEE Analysis		
Source	DF	Chi-Square	Pr > ChiSq
Density target host	1	9.14	0.0025

section, accurate results. Since Generalized Linear Models are fully parametric, they are the most powerful solution for survival analysis, even though in several statistical packages the user may find other types of analyses that are mostly non- or semi-parametric in the menu for survival analysis. However, there is – at least – one possible impediment to using Generalized Linear Models for survival analyses. Imagine a test performed to evaluate the effect of insecticide residues on survival times. While all insects in the treatment group (insecticide) are dead at day 10, 8% of the specimens in the control group are still alive at day 20, the planned end of the observation. What should be done with the data points from this 8% of the control group? Should they be left out, since no data for their longevity have been measured? This would lead to a loss of biologically meaningful data and, even more disturbing, to a bias in the interpretation, since we know that those individuals survived until at least day 20. The only thing we do not know is for how much longer they would have lived. These data points are called 'right-censored'.

A so-called log-rank test, or, more generally, a Cox regression model (= proportional hazards model), can adequately deal with censored survival data (Cox, 1972). Recently, a plethora of different studies have used such a statistical analysis for ecological investigations on insects (e.g. van Alphen *et al.*, 2003). Besides using this sort of analysis to study changes in survival time, survival analysis can also be used when it comes, e.g. to testing residence times or withdrawal times of natural enemies on patches with target and non-target hosts, or when testing the latency until a natural enemy attacks a host or prey (see van Lenteren *et al.*, Chapter 3, this volume). Briefly, the probability of dying, leaving a patch or attacking, λ, can be modified in the course of time by covariates and the Cox regression provides estimates for how the covariates, i.e. treatment effects, modify the baseline hazard of dying, leaving a patch or performing an attack. For further information, we recommend readers to consult papers that provide a thorough description of the method (e.g. Haccou and Hemerik, 1985; Haccou and Meelis, 1992; Wajnberg *et al.*, 1999; van Alphen *et al.*, 2003).

Example

Imagine a small arena no-choice test with behavioural observation of a candidate natural enemy on either target or non-target hosts (see van Lenteren *et al.*, Chapter 3, this volume for the setup). Observations are limited to one hour, after which almost all of the target hosts were attacked, and 56.7% of the non-target hosts. However, it seems that while target hosts are attacked almost immediately, the natural enemies attack non-target hosts only after a rather long period of searching the small arena, from which they cannot escape. The acceptance of non-target hosts is probably an overestimation of the host range of the natural enemy (see van Lenteren *et al.*, Chapter 3, this volume) and we thus test the latency until the host is attacked. This will elucidate whether there is a significant effect of the host species on the acceptance pattern of the natural enemy. In the Cox regression, the 43.3% of non-target hosts that remained unattacked are entered as censored observations. The Cox regression returns a highly significant ($P < 0.001$) effect of host species on the probability of being attacked. To elucidate this in detail, we plot the cumulative hazard function. This function gives the cumulated instantaneous potential for the event (i.e. the attack) to occur, given it has not yet occurred. The cumulative hazard function is thus a useful measurement of the danger of being attacked at any point in time. Here, it indicates that the probability of being attacked is 15.733 times higher per unit time for target hosts compared with non-target hosts (Fig. 13.3).

Conclusions

Conducting experiments for the assessment of non-target effects of biological control agents will be costly in terms of the man-

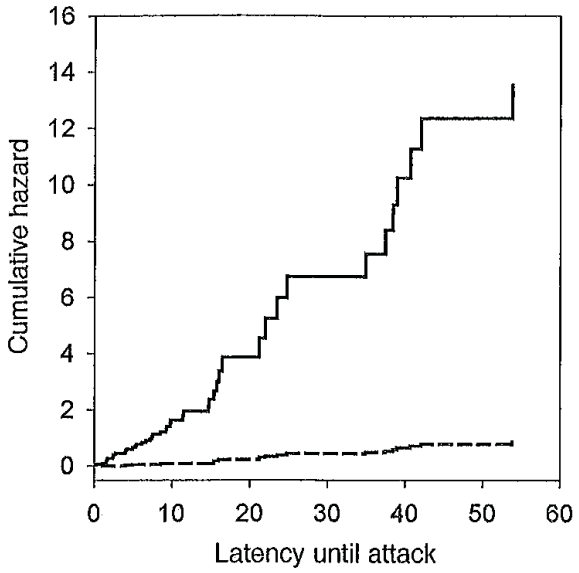

Fig. 13.3. Cumulative hazard functions for the latency until target hosts (solid line) and non-target hosts (dashed line) are attacked. Target hosts have a much higher probability per unit time of being attacked than non-target hosts.

power involved, the specimens provided for testing and the plants or plant parts needed for, e.g. host specificity tests etc. Thus, there is a high premium on using the best experimental design and the most powerful statistical methods, in order to obtain reliable test results from a reasonable amount of replicates. This is especially so, since the result we are most interested in, i.e. the probability that non-target effects do not exist, is not directly testable. What we can test is whether the null hypothesis of no effect on non-target species is wrong. If we do not find a significant effect, it very much depends upon the power of the test to decide with some confidence that no effect exists. Therefore, great care should be taken to determine the appropriate replicate number of tests. A priori power analyses, as pointed out in this chapter, are the appropriate approach here, and whenever non-significant results are stated, the power and the associated effect size should be stated in order to provide the reader with information about the degree of confidence of the results. Furthermore, the experiments should be planned in detail to ensure that no pseudoreplication occurs. Recent analyses of research papers in ecology have found a relatively high prevalence of pseudoreplication (Heffner *et al.*, 1996; Ramirez *et al.*, 2000), in spite of Hurlbert's (1984) seminal paper. Thus, the importance of avoiding pseudoreplication must be stressed here, and randomization should be used wherever possible to avoid interdependency. Fortunately, very powerful statistical techniques like Generalized Linear Models and survival analyses have become available and are now widely used in a variety of biological disciplines (e.g. Garrett *et al.*, 2004). They not only help to increase the precision of testing results but also the accuracy of tests, since they can adequately deal with non-normally distributed data that we frequently encounter in non-target effects testing. With this chapter we hope to improve the awareness of the problems, and have indicated solutions suitable for improving the quality of experiments assessing non-target effects of biological control agents.

Acknowledgements

We are grateful to B.D. Roitberg, L. Hemerik and U. Kuhlmann for reviewing an earlier version of this chapter and for their helpful comments that led to important improvements.

References

Andow, D.A. (2003) Negative and positive data, statistical power, and confidence intervals. *Environmental Biosafety Research*, 2, 75–80.

Carey, J.M. and Keough, M.J. (2002) The Variability of Estimates of Variance, and Its Effect on Power Analysis in Monitoring Design. *Environmental Monitoring and Assessment* 74, 225–241.

Cohen, J. (1998) *Statistical Power Analysis for the Behavioural Sciences.* Lawrence Erlbaum, Hillsdale, New Jersery.

Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society B* 34, 187–220.

Crawley, M.J. (1993) *GLIM for Ecologists.* Blackwell Scientific Publications, Oxford, UK.

Crawley, M.J. (2002) *Statistical Computing: An Introduction to Data Analysis Using S-Plus.* John Wiley and Sons Ltd, Chichester, UK.

Di Stefano, J. (2003) How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology* 17, 707–709.

Garcia, L.V. (2004) Escaping the Bonferroni iron claw in ecological studies. *Oikos* 105, 657–663.

Garrett, K.A., Madden, L.V., Hughes, G. and Pfender, W.F. (2004) New applications of statistical tools in plant pathology. *Phytopathology* 94, 999–1003.

Grafen, A. and Hails, R. (2002) *Modern Statistics for the Life Sciences.* Oxford University Press, Oxford, UK.

Haccou, P. and Hemerik, L. (1985) The influence of larval dispersal in the cinnabarmoth (*Tyria jacobaea*) on predation by the red wood ant (*Formica polyctena*). An analysis based on the proportional hazards model. *Journal of Animal Ecology* 54, 755–769.

Haccou, P. and Meelis, E. (1992) *Statistical Analysis of Behavioural Data. An Approach Based on Time-Structured Models.* Oxford University Press, Oxford, UK.

Heffner, R.A., Butler, M.J. and Reilly, C.K. (1996) Pseudoreplication revisited. *Ecology* 77, 2558–2562.

Hilborn, R. and Mangel, M. (1997) *The Ecological Detective. Confronting Models with Data.* Princeton University Press, Princeton, New Jersey.

Hoenig, J.M. and Heisey, D.M. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician* 55, 19–24.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression.* John Wiley and Sons Inc., New York.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54, 187–211.

Lang, A. (2004) Monitoring the impact of Bt maize on butterflies in the field: estimation of required sample size. *Environmental Biosafety Research* 3, 55–66.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models.* Chapman and Hall, New York.

Moran, M.D. (2003) Arguments for rejecting the sequential Bonferroni in eccological studies. *Oikos* 100, 403–405.

Neuhäuser, M. (2004) Testing whether any of the significant tests within a table are indeed significant. *Oikos* 106, 409–410.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S. and Hawes, C. (2003) Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology* 40, 17–31.

Peterman, R.M. (1990) Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences* 47, 2–15.

Quinn, G.P. and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists.* Cambridge University Press, Cambridge, UK.

Ramirez, C.C., Fuentes-Contreras, E., Rodriguez, L.C. and Niemeyer, H.M. (2000) Pseudoreplication and its frequency in olfactometric laboratory studies. *Journal of Chemical Ecology* 26, 1423–1431.

Rotenberry, J.T. and Wiens, J.A. (1985) Statistical power analysis and community-wide patterns. *American Naturalist* 125, 164–168.

Rothery, P., Clark, S.J. and Perry, J.N. (2003) Design of the farm-scale evaluations of genetically modified herbicide-tolerant crops. *Environmetrics* 14, 711–717.

Ruxton, G.D. and Colegrave, N. (2003) *Experimental Design for the Life Sciences.* Oxford University Press, Oxford, UK.

Steidl, R.J., Hayes, J.P. and Schauber, E. (1997) Statistical power analysis in wildlife research. *Journal of Wildlife Management* 61, 270–279.

Thomas, L. (1997) Retrospective power analysis. *Conservation Biology* 11, 276–280.

Thomas, L. and Krebs, C.J. (1997) A review of statistical power analysis software. *Bulletin of the Ecological Society of America* 78, 126–139.

van Alphen, J.J.M., Bernstein, C. and Driessen, G. (2003) Information acquisition and time allocation in insect parasitoids. *Trends in Ecology and Evolution* 18, 81–87.

Verhoeven, K.J.F., Simonsen, K.L. and McIntyre, L.M. (2005) Implementing false discovery rate control: Increasing your power. *Oikos* 108, 643–647.

Wajnberg, E., Rosi, M.C. and Colazza, S. (1999) Genetic variation in patch time allocation in a parasitic wasp. *Journal of Animal Ecology* 68, 121–133.